DOCUMENT RESUME

ED 118 618 95 TM 005 107

AUTHOR Marshall, J. Laird; Haertel, Edward H.
TITLE A Single-Administration Reliability Index for Criterion-Referenced Tests: The Mean Split-Half Coefficient of Agreement.
INSTITUTION Wisconsin Univ., Madison. Research and Development Center for Cognitive Learning.
SPONS AGENCY National Inst. of Education (DHEW), Washington, D.C.
PUB DATE [Apr 75]
CONTRACT NE-C-00-3-0065
NOTE 28p.; Paper presented at the Annual Meeting of the American Educational Research Association (Washington, D.C., March 30-April 3, 1975)

EDRS PRICE MF-$0.83 HC-$2.06 Plus Postage
DESCRIPTORS *Criterion Referenced Tests; Statistical Analysis; *Test Reliability
IDENTIFIERS Coefficient Beta

ABSTRACT

For classical, norm-referenced test reliability, Cronbach's alpha has been shown to be equal to the mean of all possible split-half Pearson product-moment correlation coefficients, adjusted by the Spearman-Brown prophecy formula. For criterion-referenced test reliability, in an analogous vein, this paper provides the rationale behind, the analysis of, computational formulas for, and characteristics of a coefficient equal to the mean of all possible split-half coefficients of agreement. In addition, the relation of this coefficient to other test indices, including those of Harris and Livingston, is presented. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. Nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

Session # 12.13

ED118618

A SINGLE-ADMINISTRATION RELIABILITY INDEX FOR CRITERION-REFERENCED TESTS: THE MEAN SPLIT-HALF COEFFICIENT OF AGREEMENT

J. Laird Marshall
and
Edward H. Haertel

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

Wisconsin Research and Cevelopment Center
for Cognitive Learning
The University of Wisconsin
Madison, Wisconsin

AERA
Washington, D.C.
April 1, 1975

Published by the Wisconsin Research and Development Center for Cognitive Learning, supported in part as a research and development center by funds from the National Institute of Education, Department of Health, Education, and Welfare. The opinions expressed herein do not necessarily reflect the position or policy of the National Institute of Education and no official endorsement by that agency should be inferred.

Center Contract No. NE-C-00-3-0065

TM005 107

In the past decade, an increased acceptance of the interrelated notions of behavioral objectives, individualized instruction and mastery learning has given rise to new kinds of educational tests. One of these new kinds of tests has as its purpose the efficient separation of the sample of examinees into two groups, often labeled "Nonmastery" and "Mastery." When there are only two courses of action available to an examinee after this kind of test is taken (i.e., stay in that instructional module, or go on to studying the next behavioral objective) these two "scores" are the only two that need be reported. There is no purpose served in further subdivision of the test scores; the dichotomy is necessary and sufficient. Such a test, composed of several items drawn from a well-defined universe, measuring a single, narrow behavioral objective, and resulting in a dichotomous classification with reference to a predetermined criterion level, is called a criterion-referenced test (CRT) in this paper.*

The differences between a CRT and the more familiar norm-referenced test (NRT) have implications for psychometric theory. One of these is the fact that the variance of the scores obtained using a CRT need not be large. Also among these are the notions that if true score is considered dichotomous, then misclassification is the primary kind of measurement error associated with a CRT, and certain other axioms on which traditional reliability is based are not satisfied. In sum, the purpose, desired score distributions, construction, outcomes, and mathematical underpinnings of

---

*It is recognized that there are a number of writers who, with varying degrees of vehemence, would disagree with this definition of a CRT. "CRT" is merely used as a label for the kind of test described above.

reliability for CRTs are not necessarily the same as for NRTs. Moreover, the meanings of reliability are, or should be, different. Whereas an NRT is reliable insofar as an examinee receives the same score on two parallel sets of data, a CRT should be reliable insofar as the examinee receives the same dichotomous categorization from the two data sets. But if classical reliability estimates are inappropriate for CRTs, what should take their place?

A number of authors (Berger, 1973; Carver, 1970; Goodman & Kruskal, 1954; Hambleton & Novick, 1973) have suggested using a rather simple dual-administration (test-retest or parallel forms) coefficient for CRT reliability. This index is frequently called the coefficient of agreement*, and is the proportion of examinees classified similarly on the two test administrations. If + and - stand for the two classifications into which the examinees are dichotomized and the following four-fold contingency table represents the results from the two test administrations:

	+	-
+	A	B
-	C	D

N

then the coefficient of agreement (here labeled P for simplicity) is

$$(1)\quad P = \frac{A + D}{N}.$$

But this coefficent is for two test administrations, requiring either a retest or parallel forms. Can this same framework (proportion of

---

* Cohen's kappa (Cohen, 1960) has also been called the "coefficient of agreement" (Swaminathan et al., 1974). The indices are related, but they are not the same.

consistent classifications) yield a CRT analogue to the familiar single administration index of internal consistency? A coefficient of agreement calculated from splitting the test into halves would be subject to the same criticism as were split-half methods with classical reliability coefficients a few decades ago, i.e., the test split chosen might yield unrepresentative results. However, a lead is suggested by the fact that the classical internal consistencey index was shown (Cronbach, 1951) to be equal to the mean of all possible split-half reliability coefficients. To make an analogy with Cronbach's alpha, then, it would seem fruitful to consider an index equal to the mean of all possible split-half coefficients of agreement. To extend the analogy further, this index is labeled coefficient beta (β).

The coefficient

There are $\binom{n}{n/2} = \nu$ possible test splits for an n-item test (where n is even) if each half is considered to be labeled (i.e., for a two-item test the split 1 / 2 is different from 2 / 1). If β is the mean of all possible split-half coefficients of agreement, then from (1),

$$\beta = \frac{1}{\nu} \sum_{s=1}^{\nu} P_s$$

$$= \frac{1}{\nu} \sum_{s=1}^{\nu} \frac{A_s + D_s}{N} ,$$

which can be rewritten

$$(2) \quad \beta = \frac{1}{N} \left[\frac{\sum_{s=1}^{\nu} (A_s + D_s)}{\nu} \right] .$$

Thus β is also the average (over persons) proportion (over test splits) of consistent classifications (+, + or -, -).

It is shown in the appendix of this paper that for any person, the proportion of test splits which yield consistent classifications is a function of that person's total score, for a given number of items and criterion level. For instance, for a 20-item test and a criterion level of 80%, a person with a score of 7 or lower will be classified consistently (nonmastery/nonmastery) on all test splits, since a score of at least 8 is needed to achieve mastery on a half-test. Likewise, a person with a score of 18 or more will be classified similarly (mastery/mastery) for all test splits. Persons with scores of from 8 to 17, however, are classified consistently for some test splits and not for others--for example, a person with a score of 12 will be classified similarly if the test split yields half-test scores of, say, 5 and 7, but not for splits which yield, say, 9 and 3.

The computing formula for coefficient β is

$$(3) \qquad \beta = \frac{1}{N}\left[\sum_{X=0}^{k-1} f_x + \sum_{X=k}^{2k-2} f_x \cdot \phi_x(X-[k-1],\ k-1) + \sum_{X=2k}^{\frac{n}{i}+k-1} f_x \cdot \phi_x(k,\ X-k) + \sum_{X=\frac{n}{i}+k}^{n} f_x\right]$$

where N = the number of persons,

X = a person's total score,

f_x = the frequency of score X in the distribution of total scores,

k = the minimum of items on either half-test that must be answered correctly to achieve a "mastery" classification on that half-test,

n = the number of items, and

$$\phi_X(a,b) = \sum_{j=a}^{b} \frac{\binom{X}{j}\binom{n-X}{n/2-j}}{\binom{n}{n/2}},$$

i.e. the proportion of splits which yield a half-test score of from a to b inclusive, given a total score of X. The derivation of this formula may be found in the appendix.

Some examples

A formula of this complexity is more suited to a computer than to hand calculation, but some examples may clarify matters. To keep computations relatively simple, the following cases are for 8 items, 10 examinees, and criterion level of 75%, yielding a cut-off score of 6 (and hence k = 3).

Example 1. Consider the total score vector $\vec{X}$ = (1,3,4,5,5,6,6,6,7,8). Note that this score distribution is unimodal, with most scores near the cut-off, and that half the examinees are classified "mastery" and half "nonmastery." This is not the kind of score distribution one would normally hope for on a CRT. Then

$$\beta = \frac{1}{10}\left[(f_0 + f_1 + f_2) + \left(f_3\phi_3(1,2) + f_4\phi_4(2,2)\right) + (f_6\phi_6(3,3) + (f_7 + f_8))\right]$$

$$= \frac{1}{10}\left[(0+1+0) + \left(1 \cdot \frac{\sum_{j=1}^{2}\binom{3}{j}\binom{8-3}{4-j}}{\binom{8}{4}} + 1 \cdot \frac{\binom{4}{2}\binom{8-4}{2}}{\binom{8}{4}}\right) + \left(3 \cdot \frac{\binom{6}{3}\binom{8-6}{1}}{\binom{8}{4}}\right) + (1+1)\right]$$

$$= \frac{1}{10}\left[1 + \left(\frac{\binom{3}{1}\binom{5}{3}}{\binom{8}{4}} + \frac{\binom{3}{2}\binom{5}{2}}{\binom{8}{4}}\right) + \frac{\binom{4}{2}\binom{4}{2}}{\binom{8}{4}} + \left(3 \cdot \frac{\binom{6}{3}\binom{2}{1}}{\binom{8}{4}}\right) + 2\right]$$

$$= \frac{1}{10}\left[1 + \frac{3 \cdot 10}{70} + \frac{3 \cdot 10}{70} + \frac{6 \cdot 6}{70} + 3 \cdot \frac{20 \cdot 2}{70} + 2\right]$$

$$= \frac{1}{10}\left[1 + .429 + .429 + .514 + 3(.571) + 2\right]$$

$$= \frac{1}{10}\ [6.085]$$

$$= .61, \text{ approximately.}$$

Example 2. $\vec{X} = (0,1,1,2,3,6,7,7,8,8)$. This score distribution is bimodal, with a gap separating the scores of the half labeled "masters" from the half labeled "nonmasters." This more closely represents the kind of score distribution one would look for when administering a test designed to separate the examinees into two groups. Here,

$$\beta = \frac{1}{10}\left[(f_0 + f_1 + f_2) + (f_3 \cdot \phi_3(1,2)) + \left(f_6 \cdot \phi_6(3,3)\right) + (f_7 + f_8)\right]$$

(note that $f_4 = f_5 = 0$)

$$= \frac{1}{10}\left[(1 + 2 + 1) + \left(1 \cdot \sum_{j=1}^{2} \frac{\binom{3}{j}\binom{5}{4-j}}{\binom{8}{4}}\right) + \frac{\binom{6}{3}\binom{2}{1}}{\binom{8}{4}} + (2 + 2)\right]$$

$$= \frac{1}{10}\left[4 + \frac{3 \cdot 10}{70} + \frac{3 \cdot 10}{70} + \frac{20 \cdot 2}{70} + 4\right]$$

$$= \frac{1}{10}\ [4 + .429 + .429 + .571 + 4]$$

$$= \frac{1}{10}\ [9.429]$$

$$= .94, \text{ approximately.}$$

Example 3. Lest one be tempted to attribute the difference in values of β to the fact that the variance of the second score distribution is about two and a half times as large as that of the first example, it should be pointed out that the magnitude of β does not rely on score variance. Thus, for $\vec{X}$ = (6,6,7,7,7,8,8,8,8,8), where all examinees are classified "masters" and where the score variance is only about a sixth as large as that of the first example, β = .91.

Adjustment for odd number of items

Thus far it has been assumed that the test has an even number of items. If n is odd, a test split is defined as resulting when one item is deleted and the remaining n-1 (even) items are divided into two sets, each containing $\frac{n-1}{2}$ items. The procedure for computing β is identical for even and odd n, except that in the latter case we first perform an additional step, for reasons explained in the appendix, replacing f_x by $\frac{(n-x)f_x + (x+1)f_{x+1}}{n}$ for X = 0,1,...,n-1 and then using n-1 in place of n in the computations of k and ϕ_x (a,b).

Properties of Coefficient β

1. Coefficient β is additive; it is the mean of its component parts. Thus each person's score makes a contribution to the value of β. Moreover, it is apparent from Equation (3) and from the analysis given in the appendix that as a score approaches the point 2k-1 (where k is the half-test cut-off score, as defined for equation (3)), it contributes successively less to the value of β; a score of 2k-1 contributes zero. If C represents the (integral) cut-off score, 2k-1 is either C or (as in this example) C-1. (See Marshall (in press) for a more thorough discussion.) What this means is that as scores depart from the cut-off, the value of β

increases, a fact that is consonant with the notion that β measures consistency of dichotomous classification.

2. Coefficient β is variance-free in the respect deemed most important by critics of a variance-dependent CRT reliability coefficient: β can take on its full range of [0, 1] even though the total score variance is zero, depending on the relative locations of the cut-off score and the (single-membered) set of test scores. It is, however, variance-dependent in other respects. As the variance approaches its maximum, coefficient beta approaches 1, which is reassuring since maximum variance obtains only when scores on an n-item test are equally divided between 0 and n, which scores indicate the clearest possible separation into masters and nonmasters. Furthermore, if β is zero, then variance is zero. These facts can be easily summarized: if variance is high, β is high; if variance is low, there is no restriction (within its range) on β.

3. For a given test type and criterion level, the value of β is not affected by the number of examinees.

4. For a given test type and criterion level, the value of β is, however, affected by the number of items: β increases as the number of items increases. A study using simulated data (Marshall, in press) indicates that although shape of score distribution also has some effect, one can prophesize reasonably well the value of β for a test twice as long via the formula $\hat{\beta}_{2n} = \frac{\beta_n(3 + \beta_n)}{2(1 + \beta_n)}$. This formula, arrived at from purely empirical grounds, is the arithmetic mean of the values obtained from the Spearman-Brown prophecy formula $f(\beta) = \frac{2\beta}{1+\beta}$ and the prediction $f(\beta) = \beta$.

5. The value of coefficient β is (usually) different for different criterion levels. There are a total of n/2 meaningful criterion levels for an n-item test, since formula (3) utilizes k, the cut-off score on a half-test. As criterion level, expressed as a fraction, approaches its meaningful limits of 2/n or 1, β generally tends toward 1, particularly for symmetric unimodal distributions.

Relations with other test indices

All results reported in this section are based on simulated data and are treated in more detail in a forthcoming report (Marshall, in press).

1. There seems to be a fairly high correlation across test types between α (i.e., KR-20) and $\overline{\beta}$, the mean value of β over criterion levels.

2. There is little if any connection between β and the index of efficiency, μ_c^2 (Harris, 1972), except that for unimodal score distributions the fluctuations of the two indices over criterion level seem to be opposite in direction.

3. For criterion levels most likely to be used in an actual test (i.e., from .6 to .9), β has a moderate correspondence with the phi coefficient, when the phi coefficient is calculated from a four-fold contingency table whose cells are the means of all possible split-half classifications, under which conditions the phi coefficient can be construed to be a single-administration index. Under these same conditions, phi is identical to Cohen's kappa when calculated from the same table. This coefficient is in turn a close lower bound to the mean of all possible split-half kappa coefficients.

4. For tests yielding unimodal score distributions, β seems to be measuring much the same thing as does k_{tx}^2 (Livingston, 1972). For

these unimodal distributions, both β and k^2_{tx} have somewhat similar ranges of values and patterns of fluctuation over criterion level. However, and just as important, this close relationship does not hold for bimodal distributions. The reason is that β is sensitive to (has minima, over criterion levels, near) the mode(s) of the score distribution, whereas k^2_{tx} is sensitive to (has minimum, over criterion levels, at) the test mean. In a unimodal distribution the mean and mode are usually proximate and the effect is the same; this is, of course, not generally the case for a bimodal score distribution. Figure 1 shows the fluctuations (over criterion level) of β and k^2_{tx} for a unimodal and a bimodal distribution, and shows the patterns described above.

Discussion

Although attention in this paper has been focused on criterion-referenced tests, it should be pointed out that coefficient beta is applicable any time that it makes sense to look at reliability as consistency of classification or consistency of decision-making based on scores from a measuring instrument, provided that the decision is based on some sort of cut-off point expressable as a percent of items responded to in a certain manner.

Second, coefficient beta can be used as a tool to help a criterion-referenced test developer search for the cutting score which best separates a population into two classifications. The procedure would be, given the test score distribution on a large, representative sample of the population, to calculate coefficient beta at all of the meaningful

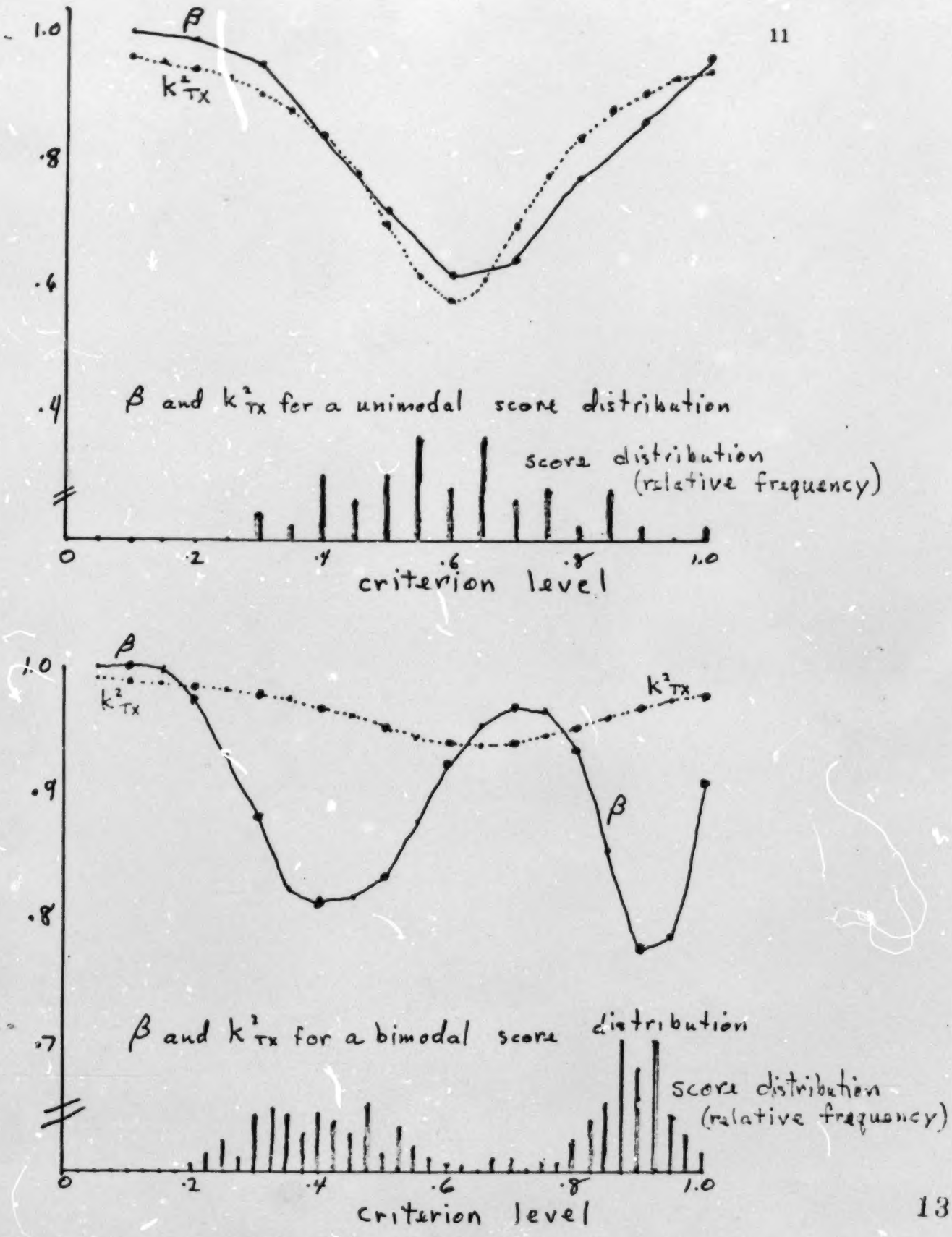

Figure 1 Fluctuations over criterion level of β and k^2_{TX}, for two selected score distributions

criterion levels which fall within a predetermined "acceptable" range (e.g. $[.7, .9]$), and then select that criterion level which yields the highest coefficient beta.

Third, mention should be made of the fact that if students respond randomly to the answers on a test, the resulting coefficient beta would not be zero, as might be expected with a traditional reliability measure. In fact, depending on the number of items, the criterion level, and the number of options per item (assuming a multiple-choice test), coefficient beta could take on a rather high value, possibly even 1. From a traditional test theory standpoint, this is disconcerting. Yet, looked at from a CRT point of view, it is understandable: for if all examinees respond randomly to a test, that is a clear indication that they are about as far from mastery as is possible; the high value of coefficient beta is an indicant that the test is classifying them as such, and reliably so. Nonetheless, a test constructer might want additional test tryout information before passing judgment about the instrument's reliability, as would be the case in the construction of a NRT.

Fourth, this paper has been concerned only with tests which result in a dichotomous classification, whereas some commercial programs prefer to have available a middle classification as well. It is shown in the appendix that coefficient beta can be extended to encompass such a trichotomous classification situation.

REFERENCES

Berger, R. J. A measure of reliability for criterion-referenced tests. Paper presented at the annual meeting of the National Council on Measurement in Education, Minneapolis, 1970.

Carver, R. P. Special problems in measuring change with psychometric devices. In Evaluative Research: Strategies and Methods. Pittsburgh: Am erican Institutes for Research, 1970.

Cohen, J. A coefficient of agreement for nominal scales. Educational and Psychological Measurement, 1960, 20, 213-220.

Cronbach, L. J. Coefficient alpha and the internal structure of tests. Psychometrika, 1951, 16, 292-334.

DMP Staff. Resource Manual, Topics 1 - 40, for Developing Mathematical Processes. Chicago: Rand McNally and Company, 1974.

Goodman, L. A., & Kruskal, W. H. Measures of association for cross classifications. Journal of the American Statistical Association, 1954, 49, 733-764.

Hambleton, R. K., & Novick, M. R. Toward an integration of theory and method for criterion-referenced tests. Journal of Educational Measurement, 1973, 10, 159-170.

Harris, C. W. An index of efficiency for fixed-length mastery tests. Paper presented at the annual meeting of the American Educational Research Association, Chicago, 1972.

Livingston, S. A. A criterion-referenced application of classical test theory. Journal of Educational Measurement, 1972, 9, 13-26.

Marshall, J.L. The mean split-half coefficient of agreement and its relation to other test indices: a study based on simulated data. Technical Report, Madison: Wisconsin Research and Development Center for Cognitive Learning, in press.

Swaminathan, H., Hambleton, R. K., & Algina, J. Reliability of criterion-referenced tests: a decision-theoretic formulation. Journal of Educational Measurement, 1974, 11, 263-267.

APPENDIX

Analysis and derivation of coefficient beta,
adjustment for tests with odd number of items,
and extension of the coefficient to trichotomous data

Definitions

Let

N = the number of examinees;

n = the number of test items;

X_p = the pth person's total score, p = 1,...,N;

c = the criterion level, expressed as a fraction $(0 < c \leq 1)$;

k = the smallest integer $\geq \frac{cn}{2}$, and hence the minimum number of items in a half-test* that must be answered correctly to receive a mastery classification on that half-test; and

X_{1p}, X_{2p} = the pth person's scores within the two half-tests, and hence $X_{1p} + X_{2p} = X_p$.

There are $\binom{n}{n/2} = \nu$ possible split-halves of the n items, if one considers each half to be labeled (i.e., for a two-item test the split 1 / 2 is different from 2 / 1). For each pair of split-halves, construct a four-fold mastery (+) / nonmastery (-) contingency table

	+	-
+	A	B
-	C	D

N

and define $P = \frac{A+D}{N}$.

---

* For now, only tests with an even number of items are considered. Tests with an odd number of items are dealt with later.

Then β is the mean of P taken over all ν possible test splits (s):

$$\beta = \frac{1}{\nu} \sum_{s=1}^{\nu} P_s$$

$$= \frac{1}{\nu} \sum_{s} \frac{A_s + D_s}{N}$$

$$= \frac{1}{N} \left(\frac{\Sigma A_s + \Sigma D_s}{\nu} \right) . \tag{A.1}$$

Analysis of the coefficient

For any given test, the set of possible scores for an individual is $\{0,1,\ldots,n\}$. For computational purposes, this is partitioned into five subsets, one or more of which may be empty for a particular n and k:

$$S_1 = \{0,\ldots,k-1\}$$

$$S_2 = \{k,\ldots,2k-2\}$$

$$S_3 = \{2k-1\}$$

$$S_4 = \{2k,\ldots,{}^{n}/_{2}+k-1\}$$

$$S_5 = \{\tfrac{n}{2}+k,\ldots,n\}.$$

(Note that k=1 implies $S_2 = \{\ \}$, and $k = \frac{n}{2}$ implies $S_4 = \{\ \}$.)

Then consider scores in each of the five subsets:

1. For $X_p \in S_1$, $X_p < k$. Thus mastery on a half-test cannot be obtained no matter how the test is split, since both X_{1p} and X_{2p} must necessarily be less than k. Hence all persons with $X_p \in S_1$ will contribute to D, as defined in the contingency table above, for all ν test splits.

2. For $X_p \in S_2$, $k \leq X_p \leq 2k-2$. Here some splits will contribute to B or C (for example, $X_p = k+1$; $X_{1p} = k$, $X_{2p} = 1$) and some will contribute to D (for example, $X_p = 2k-2$; $X_{1p} = X_{2p} = k-1$). The obvious question "Which splits?" becomes a problem of combinatorics. Since only A and D enter into Equation A.1, one need not be concerned with contributions to B and C. (These will be equally divided among B and C because of the symmetry implied in "labelling" the halves of the test.)

The question then reduces to "For a score of $X_p \in S_2$, how many D-categorizations will result?" This will happen when neither half-test is mastered, i.e. when both $X_{1p}, X_{2p} \leq k-1$.

Define $\vec{X}_{1p}$ and $\vec{X}_{2p}$ as vectors of 0's and 1's indicating incorrect/correct responses to items on each half-test. If one vector has $k-1$ 1's, the other has $X_p-(k-1)$ 1's. Moreover, since $X_p \in S_2$ and hence $X_p \leq 2k-2$, it follows that $X_p-(k-1) \leq k-1$. Thus one is interested only in those pairs of vectors in which the number of 1's in each is between these two limits, namely $X_p-(k-1) \leq$ both $X_{1p}, X_{2p} \leq k-1$. Moreover, since in the total score there are X_p 1's, there are $n-X_p$ 0's. In the half-score, if there are j 1's, there are $n/2 - j$ 0's. Thus, for $X_p \in S_2$, we can pick pairs of vectors which will yield D-categorizations in $\sum_{j=X-(k-1)}^{k-1} \binom{X}{j} \binom{n-X}{n/2-j}$ ways.

3. For $X_p \in S_3$, $X_p = 2k-1$. Thus the most "balanced" split will yield k 1's in one vector and $k-1$ 1's in the other, indicating mastery in the first case and nonmastery in the second. Other, less

2. For $X_p \in S_2$, $k \leq X_p \leq 2k-2$. Here some splits will contribute to B or C (for example, $X_p = k+1$; $X_{1p} = k$, $X_{2p} = 1$) and some will contribute to D (for example, $X_p = 2k-2$; $X_{1p} = X_{2p} = k-1$). The obvious question "Which splits?" becomes a problem of combinatorics. Since only A and D enter into Equation A.1, one need not be concerned with contributions to B and C. (These will be equally divided among B and C because of the symmetry implied in "labelling" the halves of the test.)

The question then reduces to "For a score of $X_p \in S_2$, how many D-categorizations will result?" This will happen when neither half-test is mastered, i.e. when both $X_{1p}, X_{2p} \leq k-1$.

Define $\vec{X}_{1p}$ and $\vec{X}_{2p}$ as vectors of 0's and 1's indicating incorrect/correct responses to items on each half-test. If one vector has $k-1$ 1's, the other has $X_p-(k-1)$ 1's. Moreover, since $X_p \in S_2$ and hence $X_p \leq 2k-2$, it follows that $X_p-(k-1) \leq k-1$. Thus one is interested only in those pairs of vectors in which the number of 1's in each is between these two limits, namely $X_p-(k-1) \leq$ both $X_{1p}, X_{2p} \leq k-1$. Moreover, since in the total score there are X_p 1's, there are $n-X_p$ 0's. In the half-score, if there are j 1's, there are $n/2 - j$ 0's. Thus, for $X_p \in S_2$, we can pick pairs of vectors which will yield D-categorizations in $\sum_{j=X-(k-1)}^{k-1} \binom{X}{j} \binom{n-X}{n/2-j}$ ways.

3. For $X_p \in S_3$, $X_p = 2k-1$. Thus the most "balanced" split will yield k 1's in one vector and $k-1$ 1's in the other, indicating mastery in the first case and nonmastery in the second. Other, less

"balanced" splits will yield more extreme allocations of 1's, resulting in the same mastery/nonmastery classification. Thus, for all $X_p \in S_3$, no split contributes to A or D.

4. For $X_p \in S_4$, $2k \leq X_p \leq \frac{n}{2} + k-1$. This case is similar to that of S_2. Some splits will contribute to B or C (for example, $X_p = 2k$; $X_{1p} = k+1$, $X_{2p} = k-1$) and some to A (for example, $X_p = 2k$; $X_{1p} = X_{2p} = k$). Since $X_p \geq 2k$, it cannot be that both $X_{1p}, X_{2p} < k$, and hence there are no contributions to D. Again we ignore the contributions to B and C, this time focusing attention on the contributions to A.

In this case, one needs to count those vectors such that both half-tests are mastered, i.e. where both $X_{1p}, X_{2p} \geq k$. If one half-test vector has k 1's, the other has $X_p - k$ 1's. But $X_p \in S_4$ implies $X_p \geq 2k$, which implies $k \leq X_p - k$. Thus one is interested only in those half-test vectors such that $k \leq$ both $X_{1p}, X_{2p} \leq X_p - k$. Using reasoning identical to that of case S_2, the total number of splits which will contribute to A for $X_p \in S_4$ is $\sum_{j=k}^{X-k} \binom{X}{j} \binom{n-X}{n/2-j}$.

5. For $X_p \in S_5$, $X_p \geq n/2 + k$. This says that half the items plus at least another k items are answered correctly, and thus both $X_{1p}, X_{2p} \geq k$ no matter how the test is split. Hence all ν splits contribute to A.

The coefficient

The above analysis yields an equation for β, the mean split-half coefficient of agreement. For X_p in each of the five subsets, define the following functions $\phi_i(X)$, $i=1,\ldots,5$:

1. for $0 \leq X \leq k-1$ $\quad \phi_1(X) = 1$

2. $k \leq X \leq 2k-2$ $\quad \phi_2(X) = \sum_{j=X-(k-1)}^{k-1} \binom{X}{j}\binom{n-X}{n/2-j} \Big/ \binom{n}{n/2}$

3. $X = 2k-1$ $\quad \phi_3(X) = 0$

4. $2k \leq X \leq n/2 + k-1$ $\quad \phi_4(X) = \sum_{j=k}^{X-k} \binom{X}{j}\binom{n-X}{n/2-j} \Big/ \binom{n}{n/2}$

5. $n/2+k \leq X \leq n$ $\quad \phi_5(X) = 1$

Here, $\phi_i(X)$ is the proportion of splits which contribute to A or D for a given score X.

Then Equation A.1 can be rewritten

$$\beta = \frac{1}{N} \sum_{p=1}^{N} \phi_i(X_p) \,, \tag{A.2}$$

where the index i depends on the value of X_p. Hence β has range [0,1]; it is 0 when all $X_p \in S_3$; 1 when all $X_p \in S_1 \cup S_5$.

Although Equation A.2 sums up the analysis rather simply, it is inefficient for computing purposes. A more efficient method is to generate a frequency distribution of total scores, and compute $\phi_i(X)$ only once for each possible value. In general, let f_x be the frequency of score X, $X = 0,\ldots,n$, $\sum_{X=0}^{n} f_x = N$. Then

$$\beta = \frac{1}{N} \sum_{X=0}^{n} f_x \cdot \phi_i(X),$$

where again the index i depends on the value of X.

More explicitly, since for some values of X, $\phi_1(X) = 0$ or 1,

$$\beta = \frac{1}{N}\left[\sum_{X=0}^{k-1} f_X + \sum_{X=k}^{2k-2} f_X \cdot \phi_2(X) + \sum_{X=2k}^{\frac{n}{2}+k-1} f_X \cdot \phi_4(X) + \sum_{X=\frac{n}{2}+k}^{n} f_X\right]$$

$$= \frac{1}{N}\left[\sum_{X=0}^{k-1} f_X + \sum_{X=k}^{2k-2} f_X \cdot \phi_X(X-[k-1],\ k-1) + \sum_{X=2k}^{\frac{n}{2}+k-1} f_X \cdot \phi_X(k, X-k) + \sum_{X=\frac{n}{2}+k}^{n} f_X\right]$$

where

$$\phi_X(a,b) = \sum_{j=a}^{b} \frac{\binom{X}{j}\binom{n-X}{n/2-j}}{\binom{n}{n/2}} .$$

Adjustment for odd n

For an odd number of items, a test split is defined as resulting when one item is deleted and the remaining items are divided into two sets, each containing $\frac{n-1}{2}$ items. In this case, k is the smallest integer $\geq \frac{c(n-1)}{2}$. The item deleted may be chosen in n ways, each yielding a distinct set of n-1 items to be split. Hence there are $n\binom{n-1}{(n-1)/2}$ possible split halves, if one again considers each half to be labeled.

For person p, with total score X_p, the response vector $\vec{X}_p$ has X_p 1's and $n-X_p$ 0's. Thus, for person p, X_p of the n possible choices of the deleted item will result in a set of n-1 items containing X_p-1 1's, and $n-X_p$ choices will result in a set containing

X_p 1's. Thus the contribution to coefficient β for person p, rather than $\phi_i(X_p)$, will be $\frac{X_p}{n}\phi_i(X_p-1) + \frac{n-X_p}{n}\phi_i(X_p)$ and hence, taking the mean over persons, $\beta = \frac{1}{nN}\sum_{p=1}^{N}\left[X_p \cdot \phi_i(X_p-1) + (n-X_p)\cdot\phi_i(X_p)\right]$.

As before, it is necessary to compute $\phi_i(X)$ only once for each possible value of X.

But also as before, the computation is more efficient if we utilize the frequency distribution of total scores. Recall that for a score of X_p on n (odd) items, for $n-X_p$ choices of the item deleted the total score on n-1 items will remain at X_p, and for X_p choices the total score on n-1 items will be reduced to X_p-1. The effect is that of a transformation, $\xrightarrow{t}$, on the set of total scores. In symbols,

$$X \xrightarrow{t} X \text{ in } \frac{n-X}{n} \text{ of the cases;}$$

$$X \xrightarrow{t} X-1 \text{ in } \frac{X}{n} \text{ of the cases, and hence}$$

$$X+1 \xrightarrow{t} X \text{ in } \frac{X+1}{n} \text{ of the cases.}$$

Hence, a total score of X is arrived at with frequency

$g(X) = \frac{n-X}{n} f_x + \frac{X+1}{n} f_{x+1}$. (Note that, since $f_{n+1} = 0$, $g(n) = \frac{n-n}{n} f_n + \frac{n+1}{n} f_{n+1} = 0$, and therefore $\sum_{X=0}^{n} g(X) = \sum_{X=0}^{n-1} g(X)$. Furthermore, it is easily shown that $\sum_{X=0}^{n-1} g(X) = \sum_{X=0}^{n} f_x$.) Thus, taking the mean over the transformed frequency distribution of total scores, coefficient beta is

$$\beta = \frac{1}{N} \sum_{X=0}^{n-1} g(X) \cdot \phi_i(X)$$

$$= \frac{1}{N} \sum_{X=0}^{n-1} \left[\frac{n-x}{n} f_x + \frac{x+1}{n} f_{x+1} \right] \phi_i(X) ,$$

where once again the index i depends on the value of X. Thus, in practice, the computation of β is identical for the cases of even and odd n, except that in the latter case one first performs an additional step, replacing f_x by $\frac{(n-X)f_x + (X+1)f_{x+1}}{n}$ for $X = 0,1,\ldots,n-1$ and then using $n-1$ in place of n in the computations of k and $\phi_i(X)$.

Coefficient beta and trichotomous data

The authors of some commercial instructional programs, such as Developing Mathematical Processes (DMP Staff, 1974), contend that mastery/nonmastery alone is not a sufficient categorization of test results, and that more valuable information and more appropriate teacher options become available if the test result data are trichotomized. Coefficient beta, as outlined above, is clearly not sensitive to such a trichotomization sceme.

The trichotomous coefficient of agreement in such a situation would be equal to

$$P = \frac{A+E+I}{N} ,$$

based on the table

	+	*	-
+	A	B	C
*	D	E	F
-	G	H	I

N ,

where the symbols +, *, - stand for the three categorizations. If a coefficient analogous to β were to be applicable to this sort of set-up, it should be equal to $\frac{1}{\nu}\sum_{s=1}^{\nu}\frac{A_s + E_s + I_s}{N}$, or the mean split-half trichotomous coefficient of agreement.

As it turns out, such a coefficient can be derived, although the derivation is not presented here. The analysis of this coefficient, although more complex in places than that for β, is essentially parallel to that presented earlier. Instead of partitioning the set $\{0,\ldots,n\}$ into five subsets, one partitions it into seven. Recall that for coefficient β, k is the minimum number of items on a half-test that must be answered correctly in order to receive a mastery classification. If, for trichotomized data, one in addition lets ℓ be the minimum number of items on the half-test that must be answered correctly in order to receive the middle classification, then the seven subsets of $\{0,\ldots,n\}$, together with their corresponding values of $\phi_i(X)$, $i = 1,\ldots,7$, are

$S_1 = \{0,\ldots,\ell-1\}$ $\qquad \phi_1(X) = 1$

$S_2 = \{\ell,\ldots,2\ell-2\}$ $\qquad \phi_2(X) = \sum_{j=X-(\ell-1)}^{\ell-1}\binom{X}{j}\binom{n-X}{n/2-j}\Big/\binom{n}{n/2}$

$S_3 = \{2\ell-1\}$ $\qquad \phi_3(X) = 0$

$S_4 = \{2\ell,\ldots,2k-2\}$ $\qquad \phi_4(X) = \sum_{j=u_1}^{u_2}\binom{X}{j}\binom{n-X}{n/2-j}\Big/\binom{n}{n/2}$

$S_5 = \{2k-1\}$ $\qquad \phi_5(X) = 0$

$S_6 = \{2k,\ldots,\frac{n}{2}+k-1\}$ $\qquad \phi_6(X) = \sum_{j=k}^{X-k}\binom{X}{j}\binom{n-X}{n/2-j}\Big/\binom{n}{n/2}$

$S_7 = \{\frac{n}{2}+k,\ldots,n\}$ $\qquad \phi_7(X) = 1$

where $0 < \ell < k \leq \frac{n}{2}$,

$u_1 = \max(\ell, X-[k-1])$,

and $u_2 = \min(k-1, X-\ell)$.

Note that $\ell = 1$ implies $S_2 = \{\ \}$ and $k = \frac{n}{2}$ implies $S_6 = \{\ \}$.

As before, the computation is made more efficient by utilizing the frequency distribution of total scores, and hence a formula for β_3, the mean split-half trichotomous coefficient of agreement, is

$$\beta_3 = \frac{1}{N} \sum_{X=0}^{n} f_x \cdot \phi_i(X).$$

Since $\phi_i(X)$ is 0 or 1 in four of the seven cases, this can be more explicitly rewritten

$$\beta_3 = \frac{1}{N}\left[\sum_{X=0}^{\ell-1} f_x + \sum_{X=\ell}^{2\ell-2} f_x \cdot \phi_x(X-[\ell-1],\ \ell-1) + \sum_{X=2\ell}^{2k-2} f_x \cdot \phi_x(u_1,u_2) \right.$$

$$\left. \sum_{X=2k}^{\frac{n}{2}+k-1} f_x \cdot \phi_x(k,\ X-k) + \sum_{X=\frac{n}{2}+k}^{n} f_x \right],$$

where

$$\phi_x(a,b) = \sum_{j=a}^{b} \frac{\binom{X}{j}\binom{n-X}{n/2-j}}{\binom{n}{n/2}}$$

and u_1 and u_2 are as above.

The trichotomous coefficient incorporates the same adjustments for an odd number of items as does the dichotomous coefficient, except that $n-1$ is used in calculating ℓ as well as k and $\phi_i(X)$.

Note that if the test is multiple choice, the lower of the two criterion levels should not be set near the percent of items which should be answered correctly due to chance, as this would result in unreliable classification decisions between the lower two categories. In this case, if there are a significant number of nonmasters in the population, the value of β_3 would tend to be rather low, as would be expected.